**Backstory**

The pandemic and a hiring freeze have led to a shortage of train operators, conductors and workers in New York City, forcing thousands of subway trips to be canceled. In February 2019, Thousands of subway trips in New York City have been canceled because the pandemic and a related hiring freeze have battered the work force and left a shortage of train operators, conductors and workers.

It is very important to know the number of canceled flights and the reasons for cancellation to avoid continued loss.

My work in this project is to clean, explore, aggregate, and visualize the data of same month previous years to Comparing the number of trips. Therefore, from MTA turnstile data I will use the data of February 2018, 2017, and 2016 and my focus will be mainly on UNIT, STATION, DATE, TIME, ENTRIES and EXITS columns from the data.

**Goal**

The goal of this project is preprocessing and analyzing MTA Turnstile dataset to be used in Determining the station with the most canceled trips and determining the number of cancellations compared to previous years.

**Need/ Question**

What are the dates on which the largest number of cancellations occurred?

What is the reason for cancellation?

Is hiring freeze one of the reasons?

Who benefits from exploring this question?

NYC MTA

**Data Description:**

What dataset(s) do you plan to use, and how will you obtain the data?

* The data that will be used are: MTA turnstile data:

February 2016, February 2017, and February 2018

* What characteristics/features do you expect to work with?

Exploratory Data Analysis